1 引言

近年来，人们逐渐从信息化时代迈向了数据时代，各种数据爆炸式地增长，数据消费也在日益增多，大量的信息、知识和利润隐藏在这些数据中。如何更有效地利用这些数据，已经成为这个时代下人们共同探索的问题之一。

在这次大作业中，我将对 Adult 数据集进行全面的分析：首先探索数据集中各特征的分布信息；再划分数据集，尝试多种分类模型；最后比较这些模型在 Adult 数据集上的预测结果（分析代码均基于 Python 语言，相关工具和库包可参见附录 **??**）。

2 探索 Adult 数据集

2.1 Adult 数据集的基本信息

Adult 数据集 [1] 也称人口普查收入（Census Income）数据集，来源于美国 1994 年的人口普查数据库，可以作为二分类数据集，用来预测居民年收入是否超过 50K$，其基本信息可参见表 1。

表 1: Adult 数据集的基本信息

属性	值	属性	值
数据集特征	多变量	相关应用	分类
实例数	48842	捐赠日期	1996.5.1
领域	社会	是否有缺失值	有
属性特征	类别型或整数	官网访问次数	1188850
属性数目	14		

Adult 数据集的每个实例包含 14 个属性，其含义、数据类型、取值范围等基本信息见表 2。

表 2: Adult 数据集的基本信息

特征名	含义	数据类型	类别数
age	年龄	整数	-
workclass	工作类型	类别型	8
fnlwgt	序号	整数	-
education	教育程度	类别型	16
education-num	受教育时间	整数	-
marital-status	婚姻状况	类别型	7
occupation	职业	类别型	14
relationship	家庭关系	类别型	6
race	种族	类别型	5
sex	性别	类别型	2
capital-gain	资本收益	整数	-
capital-loss	资本损失	整数	-
hours-per-week	每周工作小时数	整数	-
native-country	原籍	类别型	41

2.2 数据预处理

我首先使用 pandas 库读取 Adult 数据集，将其存储为 pandas 库中的 DataFrame 格式，随机打印出其中几个实例，对该数据集进行初步的观察，结果如下。

```
 age  work_class  fnlwgt  education  education_num       marital_status
  24     Private  269799  Assoc-voc             11        Never-married
  35           ?  169809  Bachelors             13  Married-civ-spouse
  51     Private  257126       10th              6  Married-civ-spouse
  72     Private  107814    Masters             14        Never-married
  33     Private  205950    HS-grad              9        Never-married

      occupation     relationship     race     sex  capital_gain
Exec-managerial   Not-in-family    White    Male             0
              ?         Husband    White    Male             0
   Craft-repair         Husband    White    Male             0
 Prof-specialty   Not-in-family    White    Male          2329
  Other-service       Own-child    White    Male             0

 capital_loss  hours_per_week  native_country  income
            0              40   United-States  <=50K.
            0              20   United-States   <=50K
            0              40   United-States  <=50K.
            0              60   United-States   <=50K
            0              40   United-States   <=50K
```

从以上的初步观察可以得知，Adult 数据集存在数据缺失的情况（如第 3 行和 10 行的“?”），我对整个数据集进行统计后，发现数据集中共有 3620 个实例存在缺失值，而其中 2799 个实例的缺失值多于 1 个（表 3）。

表 3: Adult 数据集缺失值分布

	无缺失值	缺失 1 个特征	缺失 2 个特征	缺失 3 个特征
实例数	45222	821	2753	46

考虑到数据集中存在缺失值的实例数较少（仅占总数的 7.41%），且缺失的均为类别型变量，若用一般的方式填补会带来较大的偏差，因此我选择将这些实例直接删除，清理缺失值后的 Adult 数据集包含 45222 个实例，且各类别型特征的类数并未因此受到影响（如 native-country 特征原本包含 41 类，处理后依然包含 41 类）。

2.3 Adult 数据集中各特征的分布

对 Adult 数据集进行检查和清理后，我开始探索 Adult 数据集中各特征的分布，我将数值型特征和类别型特征分别处理：对于数值型特征，我主要关注其数字特征（如均值，方差等）及分布密度；对于类别型特征，我主要关注其具体的分布情况。

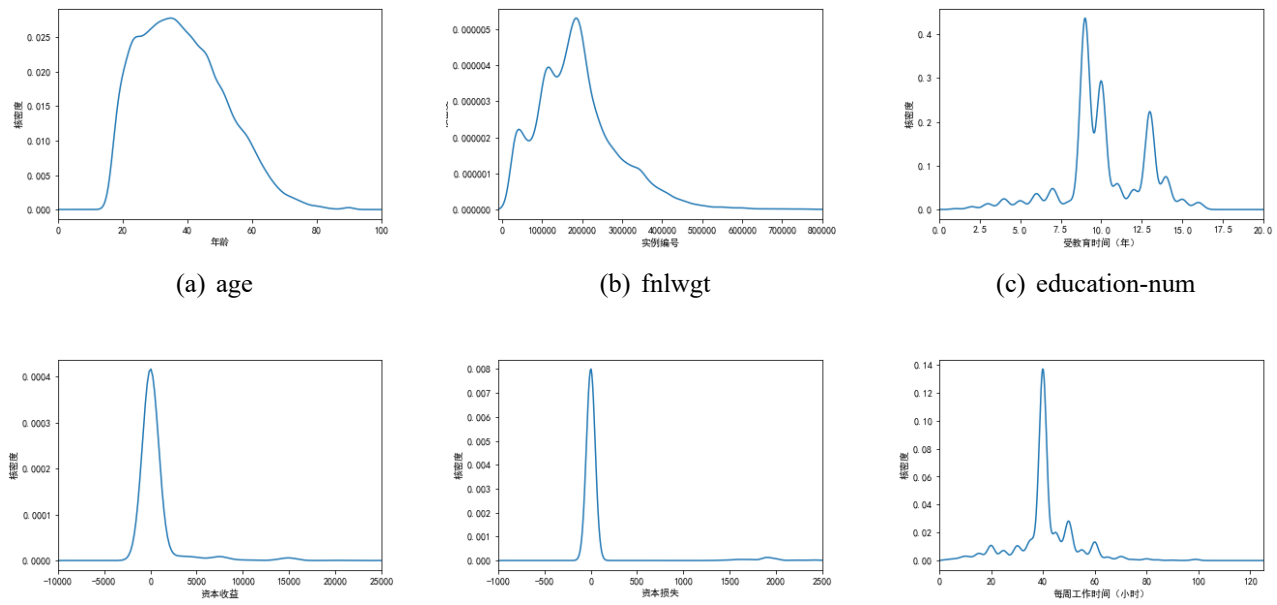
2.3.1 数值型特征的分布

Adult 数据集中的数值型特征为：age, fnlwgt, education-num, capital-gain, capital-loss 以及 hours-per-week。我首先统计其均值、标准差等数字特征，相应结果如表 4所示。

表 4: Adult 数据集数值型特征的数字特征

特征名	均值	标准差	最大值	最小值	上四分位数	下四分位数
age	38.548	13.218	90.000	17.000	47.000	28.000
fnlwgt	18976.470	10563.920	1490400.000	13492.000	237926.000	117388.200
education-num	10.118	2.553	16.000	1.000	13.000	9.000
capital-gain	1101.430	7506.430	99999.000	0.000	0.000	0.000
capital-loss	88.595	404.956	4356.000	0.000	0.000	0.000
hours-per-week	40.938	12.008	99.000	1.000	45.000	40.000

为探索各数值型特征的分布趋势，我作出了相应的概率核密度分布图（高斯核），见图 1。



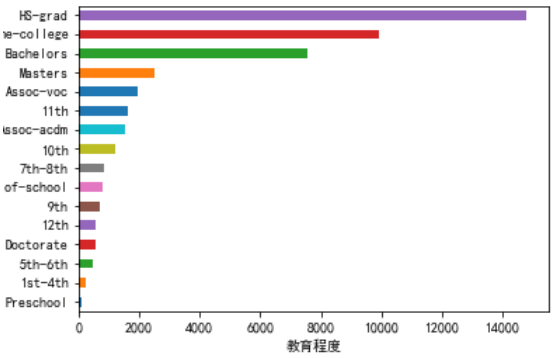
(a) age (b) fnlwgt (c) education-num

(d) capital-gain (e) capital-loss (f) hours-per-week

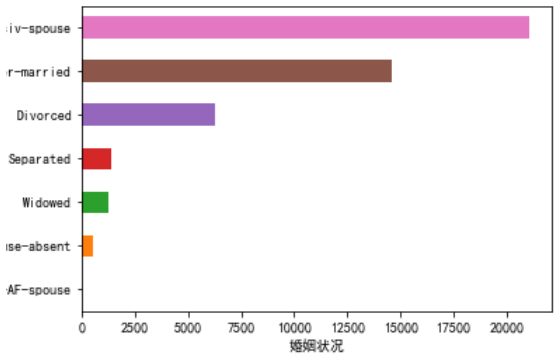
图 1: Adult 数据集数值型特征概率核密度分布

2.3.2 类别型特征的分布

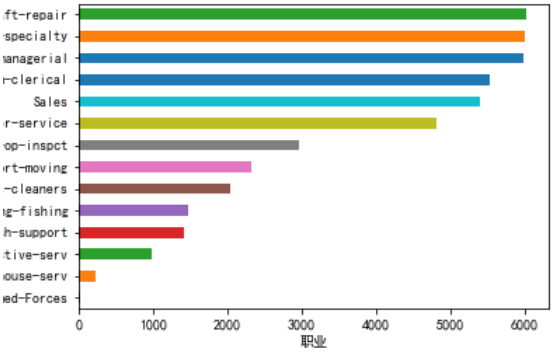
Adult 数据集中的类别型特征包含：workclass, education, marital-status, occupation, relationship, race, sex 以及 native-country。我将其分布表示为条形图或饼图（图 2）。



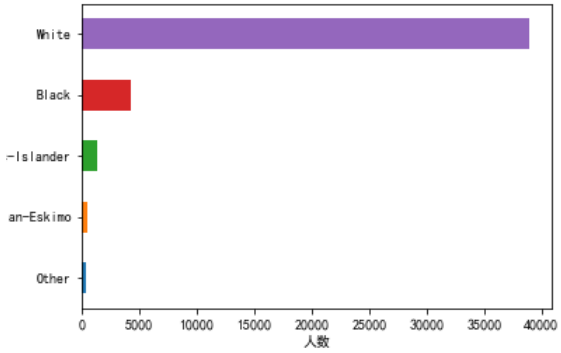
(a) education



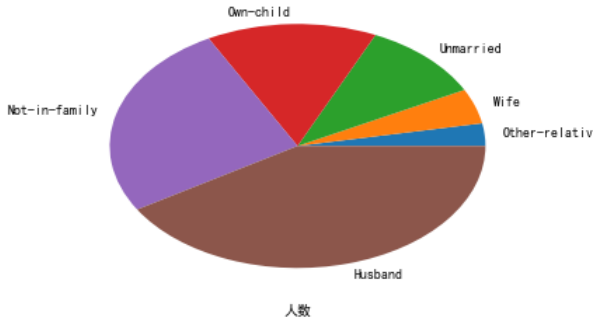
(b) marital-status



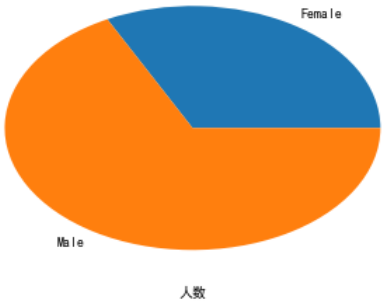
(c) occupation



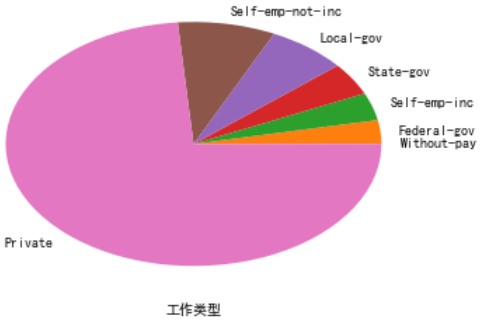
(d) race



(e) relationship



(f) sex



(g) workclass

图 2: Adult 数据集数值型特征概率核密度分布

A 附录

A.1 作业中使用的工具及库包

参考文献

[1] D. Dheeru and E. Karra Taniskidou, “UCI machine learning repository,” 2017.